Initialization Methods

Thang Vu

Stochastic Gradient Descent

- Gradient Descent:

$$\theta_i \leftarrow \theta_{i-1} - \eta \nabla C(\theta_{i-1})$$

  - In which

$$\nabla C(\theta_{i-1}) = \frac{1}{R} \sum_r \nabla C^r(\theta)$$

- Stochastic Gradient Descent:
  - Pick an example x^r

$$\theta_i \leftarrow \theta_{i-1} - \eta \nabla C^r(\theta_{i-1})$$

Stochastic Gradient Descent

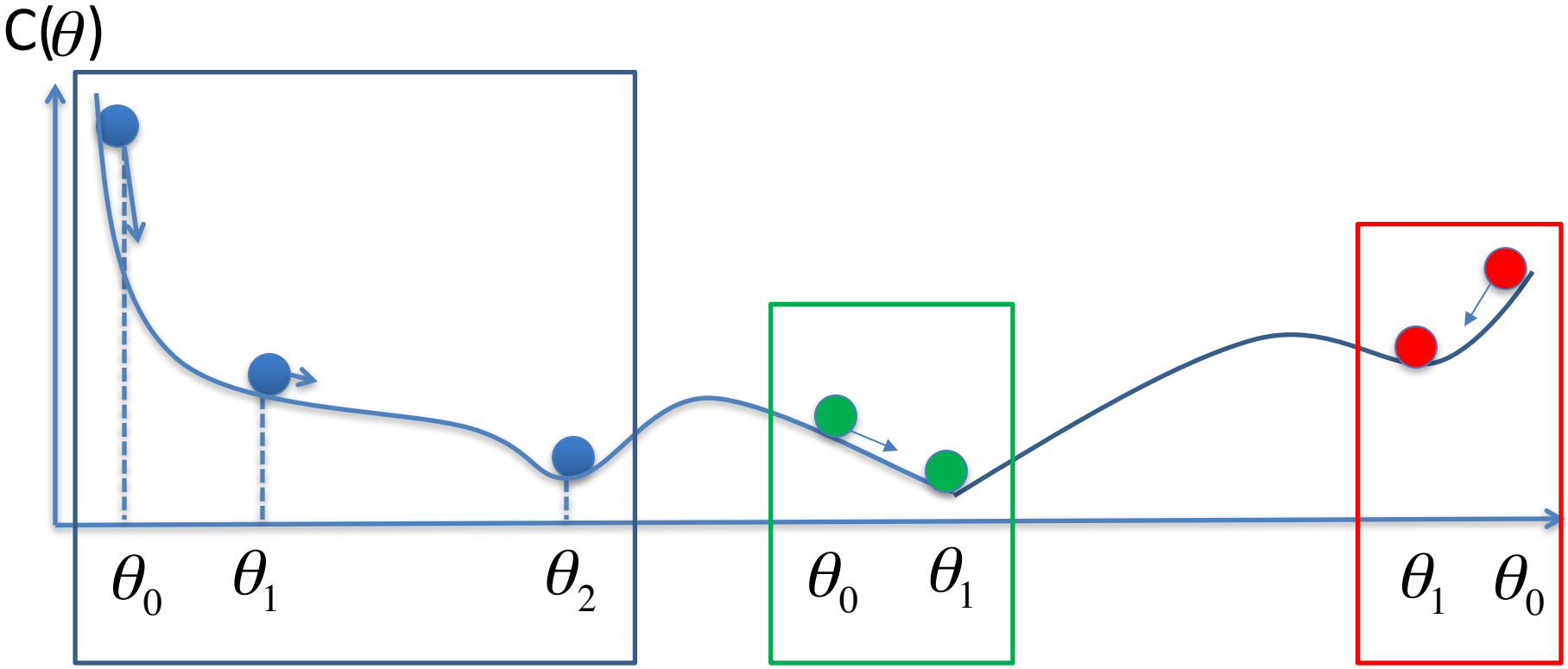
- Mini-batch Gradient Descent:
  - Pick B examples as a batch b
  - B is the batch size

$$\theta_i \leftarrow \theta_{i-1} - \eta \frac{1}{B} \sum_{x_r \in b} \nabla C^r(\theta_{i-1})$$

- Mini-batch Gradient Descent is faster than Stochastic Gradient Descent
  - Less updates
  - Better parallelization
- Important: Shuffle the data after each epoch

Challenges of Gradient Descent

- Depending on the intialization points, we will obtain different models and therefore different results



Initialization Methods

- Uniform Distribution
- Normal Distribution
- Xavier or Glorot methods - 2010
  - Either using uniform or normal distribution
- Kaiming or He methods – 2015
  - Either using uniform or normal distribution

Uniform or Normal Distribution

- Uniform
  - Values are drawn from a uniform distribution U(a,b)
  - a is lower bound, e.g. 0 and
    b is the upper bound, e.g. 1
- Normal
  - Values are drawn from a normal distribution
    $N(mean, std^2)$
  - mean is the mean value, e.g. 0
    and std is the standard deviation, e.g. 1

Xavier or Glorot Methods

- Understanding the difficulty of training deep feedforward neural networks, Xavier Glorot and Yoshua Bengio, 2010
- Uniform:
  - Values are drawn from a uniform distribution U(-a,a)

$$a = gain \cdot \sqrt{\frac{6}{fan_{in} + fan_{out}}}$$

- Normal:
  - Values are drawn from a normal distribution N(0, std^2)

$$std = gain \cdot \sqrt{\frac{2}{fan_{in} + fan_{out}}}$$

Kaiming or He Methods

- Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Kaiming He et al 2015
- Uniform:
  - Values are drawn from a uniform distribution U(-b,b)

$$\mathrm{b} = gain \,.\, \sqrt{\frac{3}{fan_{mode}}}$$

mode could be either in or out

- Normal:
  - Values are drawn from a normal distribution N(0, std^2)

$$std \;= gain \,.\, \sqrt{\frac{2}{fan_{mode}}}$$

Kaiming or He Methods

- Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Kaiming He et al 2015

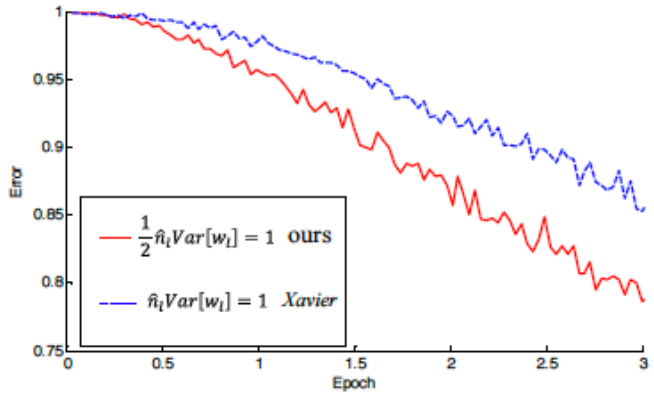


Figure 2. The convergence of a **22-layer** large model (B in Table 3). The x-axis is the number of training epochs. The y-axis is the top-1 error of 3,000 random val samples, evaluated on the center crop. We use ReLU as the activation for both cases. Both our initialization (red) and "*Xavier*" (blue) [7] lead to convergence, but ours starts reducing error earlier.

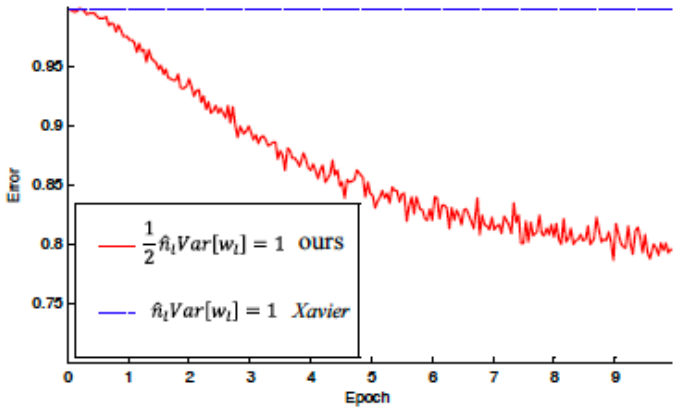


Figure 3. The convergence of a **30-layer** small model (see the main text). We use ReLU as the activation for both cases. Our initialization (red) is able to make it converge. But "*Xavier*" (blue) [7] completely stalls - we also verify that its gradients are all diminishing. It does not converge even given more epochs.

Why Kaiming or He Methods?

- We want that the variance of the input is equal to the variance of the output independent from the layer

$$Var(z^l) = Var(x)$$

Why Kaiming or He Methods?

$$
\begin{aligned}
Var(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2 + Y^2 + 2XY] - (E[X] + E[Y])^2 \\
&= (E[X^2] + E[Y^2] + 2E[XY]) - ((E[X])^2 + (E[Y])^2 + 2E[X]E[Y]) \\
&= (E[X^2] + E[Y^2] + 2E[X]E[Y]) - ((E[X])^2 + (E[Y])^2 + 2E[X]E[Y]) \\
&= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 \\
&= Var(X) + Var(Y)
\end{aligned}
$$

$$
\begin{aligned}
Var(XY) &= E[(XY)^2] - (E[XY])^2 \\
&= E[X^2]E[Y^2] - (E[X]E[Y])^2 \\
&= (Var(X) + (E[X])^2)(Var(Y) + (E[Y])^2) - (E[X])^2(E[Y])^2 \\
&= Var(X)Var(Y) + (E[X])^2 Var(Y) + Var(X)(E[Y])^2
\end{aligned}
$$

Why Kaiming or He Methods?

- $z_i = W_{i,1}\, a_1 + W_{i,2}\, a_2 + \ \ldots + W_{i,n}\, a_n + b_i$
- $Var(z_i) = Var\left(W_{i,1}\, a_1 + \cdots W_{i,n}\, a_n + \ b_i\right)$

$= n\ * Var\left(W_{i,j}\, a_j\right)$

$= n\ * Var\left(W_{i,j}\right) Var\left(a_j\right) + E\left(W_{i,j}\right)^2 Var\left(a_j\right)$

$+ Var\left(W_{i,j}\right) E\left(a_j\right)^2$

$= n\ * Var\left(W_{i,j}\right) Var\left(a_j\right) + Var\left(W_{i,j}\right) E\left(a_j\right)^2$

$= n * \left(Var\left(W_{i,j}\right)\ * \left(Var\left(a_j\right) + E\left(a_j\right)^2\right)\right.$

$= \ n * Var\left(W_{i,j}\right) * E(a_j^2)$

Why Kaiming or He Methods?

- $\mathrm{E}(a^2) = \int_{-\infty}^{\infty} a^2 P(a) da$

$$= \int_{-\infty}^{\infty} \max(0, z)^2 P(z) dz$$ ⟸ **ReLU**

$$= \int_{0}^{\infty} \mathrm{z}^2 P(z) dz$$

$$= 0.5 * \int_{-\infty}^{\infty} \mathrm{z}^2 P(z) dz$$

$$= 0.5 * E(z^2) = 0.5 * Var(z)$$

$Because\ E(z) = 0$

Why Kaiming or He Methods?

- $Var\left(z_j^l\right) = 0.5 \ * \ n * Var\left(W_{i,j}^l\right) * Var\left(z_j^{l-1}\right)$
- $Var\left(z^l\right) = 0.5 \ * \ n * Var\left(W^l\right) * Var\left(z^{l-1}\right)$

$$Var\left(z^l\right) = Var(x)\left(\prod_{l=2}^{l}\frac{n^l}{2}Var\left(W^l\right)\right)$$

$Var\left(z^l\right) = Var(x)$ only if

$\prod_{l=2}^{l}\frac{n^l}{2}Var\left(W^l\right) = 1$, i.e. $\frac{n^l}{2}\mathrm{Var}\left(W^l\right) = 1$